German International University of Applied Sciences
Informatics and Computer Science
Dr. Nada Sharaf
TA. Mariam Ali
TA. Omaima Ahmed

Big Data & NoSQL Databases, Spring 2025
Project
Due date is Thursday, May 22nd, 2025 at 11:59 PM
Submitted in groups of 3-5 (can be cross tutorial)
NO exceptions

You are working as a data analyst for a company optimizing urban mobility solutions. You've been provided with a dataset containing detailed information about yellow taxi trips in New York City, including trip durations, distances, pickup and drop-off locations, and fare components. Your role is to explore, clean, and analyze this data to generate insights and develop models that support city-wide decision making for drivers and riders.

Requirements

**** In this project you are allowed to use any of the tools, frameworks and technologies covered throughout the course(Cassandra, MongoDB, Spark, etc..)***

1. Data Cleaning & Engineering:

a. Begin by exploring both datasets: taxitripdata.csv and taxizonegeo.csv. You are expected to clean and prepare the data in a way that allows for reliable querying and analysis.
b. Some values may be missing explicitly (e.g., NULL) or implicitly (e.g., zeroes or empty strings). Handle such cases thoughtfully.
c. Drop duplicates and any columns that don't contribute meaningfully to your analysis .
d. You are expected to create new columns that provide useful insights:
   i. Calculate the **trip duration** in minutes and add it as a new column.
   ii. Compute the **total trip cost** using: fare_amount , extra, mta_tax, tip_amount, tolls_amount
e. Your final dataset should combine relevant information from both files. Make sure the zone-based location data from taxizonegeo.csv is meaningfully integrated with the trip data to enrich your analysis.

2. Analytical Queries:

Write the following queries that provide different insights into the data. You may implement them using **Spark SQL** or **DataFrame API**. Example queries include:

a) What is the most common payment type used per **time of day** (morning, afternoon, evening)?

b) Which boroughs generate the highest total revenue based on pickup locations, and how do they compare in terms of trip volume?

c) What is the average **tip amount per passenger count**?

d) What are the best 5 locations for drivers to pick up passengers from and at which time of the day?

e) What are the top 5 longest trips recorded in the dataset and display their corresponding trip duration, fare, pickup and dropoff zones, and payment type. Comment on whether these trips also resulted in high fares or if any anomalies are observed.

f) Which pickup and drop-off borough combinations represent the most frequent inter-borough travel flows? Present the top routes by trip count, and optionally include revenue per route.

3. SparkML Task: Trip Profiling: Predict Likelihood of High Tipping

- Use SparkML to build **3 machine learning models** to predict whether a trip is likely to result in a high tip, based on its characteristics (e.g., time, distance, passenger count). The focus here is on profiling the trip itself, not the rider.
- To prepare:
    - Create a binary target column:
        - high_tip = 1 if tip_amount > 0.15 × fare_amount
        - high_tip = 0 otherwise
    - Bin or normalize any necessary columns, and select relevant features such as: passenger_count, trip_distance, trip_duration_minutes, pickup_hour or time_of_day, fare_amount
    - Other engineered features (e.g., fare_per_mile, pickup/dropoff borough if encoded)
    - Combine all selected features into a single vector using VectorAssembler
- Split your data into training and testing sets, and evaluate the accuracy of each model.
- Try at least **three different classifiers** (e.g., Logistic Regression, Decision Tree, Random Forest).
- Summarize your findings and comment on which model performs best and why. You may also highlight the most important features driving the prediction (if supported by the model)

Data Description

Column	Type	Nullable	Description
vendor_id	text	required	A code indicating the TPEP provider that provided the record. 1= Creative Mobile Technologies, LLC; 2= VeriFone Inc
pickup_datetime	datetime	nullable	The date and time when the meter was engaged.
dropoff_datetime	datetime	nullable	The date and time when the meter was disengaged.
passenger_count	integer	nullable	The number of passengers in the vehicle. This is a driver-entered value
trip_distance	numeric	nullable	The elapsed trip distance in miles reported by the taximeter.
rate_code	string	nullable	The final rate code in effect at the end of the trip. 1= Standard rate 2=JFK 3=Newark 4=Nassau or Westchester 5=Negotiated fare 6=Group ride
store_and_fwd_flag	string	nullable	This flag indicates whether the trip record was held in vehicle memory before sending it to the vendor, aka “store and forward,” because the vehicle did not have a connection to the server. Y= store and forward trip N= not a store and forward trip
payment_type	string	nullable	A numeric code signifying how the passenger paid for the trip. 1= Credit card 2= Cash 3= No charge 4= Dispute 5= Unknown 6= Voided trip

Column	Type	Nullable	Description
fare_amount	numeric	nullable	The time-and-distance fare calculated by the meter
extra	numeric	nullable	Miscellaneous extras and surcharges. Currently, this only includes the \$0.50 and \$1 rush hour and overnight charges.
mta_tax	numeric	nullable	\$0.50 MTA tax that is automatically triggered based on the metered rate in use
tip_amount	numeric	nullable	Tip amount – This field is automatically populated for credit card tips. Cash tips are not included
tolls_amount	numeric	nullable	Total amount of all tolls paid in the trip.
imp_surcharge	numeric	nullable	$0.30 improvement surcharge assessed trips at the flag drop. The improvement surcharge began being levied in 2015.
total_amount	numeric	nullable	The total amount charged to passengers. Does not include cash tips
pickup_location_id	string	nullable	TLC Taxi Zone in which the taximeter was engaged
dropoff_location_id	string	nullable	TLC Taxi Zone in which the taximeter was disengaged

Deliverables

- Your Source Code: Your program should implement the specifications indicated above.
- Part of the grade will be on how readable your code is. Use explanatory comments whenever possible
- Please submit your work by compressing it into a zip file and sending it to **bigdata602.25@gmail.com** with subject: “Project” and IDs of the team members in the body of the email).
- You will be evaluated based on the submitted notebook **before the deadline only.**

Note: You will be asked for the reasoning of any actions, queries or decisions you have taken in the implementation of this project during the evaluations that have led to your answers/results